Biometrika (2001), **88,** 4, *pp.* 933–948

Printed in Great Britain

Inference for covariates that accounts for ascertainment and random genetic effects in family studies

BY RUTH M. PFEIFFER, MITCHELL H. GAIL
National Cancer Institute, Bethesda, Maryland 20892, U.S.A.
pfeiffer@mail.nih.gov gailm@mail.nih.gov

AND DAVID PEE
Information Management Services Inc., Rockville, Maryland 20852, U.S.A.
pee@rvims.nci.nih.gov

SUMMARY

Family studies to identify disease-related genes often collect families with multiple cases. If environmental exposures or other measured covariates are also important, they should be incorporated into these genetic analyses to control for confounding and increase statistical power. We propose a two-level mixed effects model that allows us to estimate environmental effects while accounting for varying genetic correlations among family members and adjusting for ascertainment by conditioning on the number of cases in the family. We describe a conditional maximum likelihood analysis based on this model. When genetic effects are negligible, this conditional likelihood reduces to standard conditional logistic regression. We show that the simpler conditional logistic regression typically yields biased estimators of exposure effects, and we describe conditions under which the conditional logistic approach has little or no bias.

Some key words: Ascertainment correction; Conditional logistic regression; Correlated binary data; Family study; Misspecified model; Nested random effects model.

1. INTRODUCTION

This paper was motivated by data on 150 families that were chosen to study the genetic aspects of nasopharyngeal carcinoma in a collaborative study supported by the National Institutes of Health in Bethesda, Maryland, and the Institute of Epidemiology at the National Taiwan University in Taipei. Families were included only if they had two or more affected family members. Although the primary purpose of this study is to detect genes that predispose to nasopharyngeal carcinoma, it is important to determine if environmental covariates also play a role. If so, these environmental factors should be incorporated into the genetic analysis to control for confounding and improve the power to find genes associated with the carcinoma. It is thus desirable to be able to use data from family studies to estimate the effects of environmental exposures before attempting to identify susceptibility genes. This is challenging because (i) the families were selected only if they had multiple affected members and (ii) unmeasured familial and genetic effects that induce correlated responses need to be taken into account. A natural approach to account for ascertaining families with a fixed number of cases would be to conduct a matched case-

control analysis with matching on family and to use conditional logistic regression that conditions on the number of cases in the family. We show that this approach can lead to underestimates of exposure effects if genetic correlations are ignored.

In § 2 we introduce a two-level mixed effects model to account for common familial effects and for different genetic correlations among family members. We adjust for ascertainment by conditioning on the number of cases in the family and performing conditional maximum likelihood analysis based on this mixed effects model. We show in § 3 that estimators based on this model yield consistent estimators of covariate effects when the covariate has no effect on disease status even with a misspecified random effects distribution. A special case of misspecification results in standard conditional logistic regression. We show that estimators based on standard conditional logistic regression yield consistent estimators of covariate effects in three situations: when the genetic influences are negligible; when the disease is rare and the familial and individual random genetic and fixed effects are small; and when the covariate has no effect on disease status. In all other situations estimators of covariate effects based on standard conditional logistic regression are biased. An approximation to the bias is given for the case of small covariate effects. For the special case of families of size two, such as a case-control study of pairs of siblings, we derive explicitly the bias of the conditional logistic regression estimator of the covariate effect. The results in § 3 do not require any assumptions on the distributions of the familial and individual genetic random effects. In § 4, we specialise the distribution of the genetic random effects to reflect Fisher's (1918) model for polygenic effects. We compare the performance of the two-level mixed effects model with that of conditional logistic regression using data simulated from the polygenic model. The conditional logistic regression bias estimates from the simulations are compared with analytical bias calculations. In § 5 we analyse a subset of the nasopharyngeal carcinoma data using the different analytical approaches. In § 6 we discuss our results and related work and mention other applications of the two-level mixed effects model, including applications to longitudinal studies.

2. The random effects model and estimation method

2·1. *Model formulation*

The family data consist of a binary disease status variable Y_{ij} for the jth member of the ith family, together with a covariate variable X_{ij}, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. By a family we mean a group of people related by blood or marriage. To avoid indefinite extension of blood relationships throughout a population, we confine attention to families consisting of a current generation of individuals of reproductive age, their siblings and offspring and at most two generations of ancestral relatives. We model the probability distribution of the disease status Y_{ij} as a function of the covariate X_{ij}, the random familial effect a_i, which affects all family members equally, and an individual level random genetic effect g_{ij} for the jth individual in the ith family:

$$\text{logit}(p_{ij}) = \text{logit pr}(Y_{ij} = 1 \mid a_i, g_{ij}, X_{ij}) = \mu + a_i + g_{ij} + \beta X_{ij}. \tag{2·1}$$

The a_i and g_{ij} have expectation zero, and $(a_i, g_i) = (a_i, g_{i1}, \ldots, g_{in_i})$ are assumed to be independent and distributed with joint distribution $F(a, g; \lambda)$ where λ denotes a finite set of parameters. The g_{ij}'s are correlated within the ith family. Different families with different pedigree structures have different distributions $F(a, g; \lambda)$, but these distributions are assumed to depend on the same vector of parameters λ. The covariate vectors X_{ij}

are assumed to be centred, independent and identically distributed for all i and j and independent of a_i and g_{ij}.

In the application to the carcinoma data, we assume that the a_i and g_{ij} are normally distributed and that the covariance of the g_{ij}'s corresponds to an additive genetic variance (Fisher, 1918) as described in § 4·2. We also assume that the a_i are independent of the g_{ij}. The results in § 3, however, hold for general joint distributions of the random effects.

Model (2·1) is appealing because it allows one to combine information on an individual's measured characteristics and covariates, X_{ij}, and on the genetic liability g_{ij}. In this model β describes the increase in log relative odds from a unit increase in exposure, X, for an individual conditional on the random effects. In the nasopharyngeal carcinoma study, the scientists plan to measure candidate genes. Once the genes are measured, they can be included as known covariates in model (2·1). The parameter β for an environmental effect in the model with measured genes would correspond more closely to the β in model (2·1) than to the parameter β^* in a 'population averaged' model (Zeger et al., 1988) of the type $\text{logit}\{\text{pr}(Y_{ij}=1|X_{ij})\}=\mu^*+\beta^* X_{ij}$. We are primarily interested in the subject-specific parameter β, rather than the population-averaged β^* as discussed by Zeger et al.

Model (2·1) is an extension of the widely used random effects model that allows for a cluster-specific intercept a_i, but assumes that the Y_{ij}'s are conditionally independent, given a_i and the measured X_{ij}; see for example Neuhaus et al. (1991). Diggle et al. (1994, Ch. 9) presents similar generalised linear mixed models for exponential families with canonical links. Linear mixed models are often used to study quantitative genetic traits; see for example Amos (1993). For dichotomous traits, model (2·1) reduces to a model presented by Houwing-Duistermaat & van Houwelingen (1998) if a_i and the covariates X_{ij} are omitted, and specialises to a model used by Burton et al. (2000) when a_i and g_{ij} are normally distributed. Witte et al. (1999) used a fixed effects logistic model in which g_{ij} were measured, rather than random effects, to analyse family studies.

Under the logistic model (2·1), the marginal probability of the response in the ith family can be written as

$$\text{pr}(Y_{i1},\ldots,Y_{in_i}|X_{i1},\ldots,X_{in_i})=\int\ldots\int\prod_{j=1}^{n_i}p_{ij}^{y_{ij}}q_{ij}^{1-y_{ij}}\,dF(a,g), \tag{2·2}$$

where $q_{ij}=1-p_{ij}$.

2·2. *The ascertainment correction*

To account for the fact that selected families have specified numbers of cases, the likelihood function of the data should be conditioned on the ascertainment event. A family was included in the carcinoma study only if there were at least two affected family members. As nasopharyngeal carcinoma is a rare disease, most of the families had exactly two diseased members. In order to simplify the calculations we condition not on the event $\sum_j Y_{ij}\geqslant 2$, but on the exact number of affected family members, $\sum_j Y_{ij}=k_i$, for $k_i\geqslant 2$. We derive analytical results for the case of $k_i=2$; they are easily generalised for $k_i=3,4,\ldots,$ and contributions to the loglikelihood can be added from families with different numbers of affected family members.

For simplicity we rearrange the observed data from each family so that the two diseased family members are Y_{i1} and Y_{i2}. To avoid complex notation, we let

$$d_i(\beta)=\prod_{j=1}^{n_i}\{1+\exp(\mu+a_i+\beta X_{ij}+g_{ij})\}^{-1}. \tag{2·3}$$

The conditional distribution for family i can be written as

$$\begin{aligned}\mathrm{pr}\left(Y_{i1},\ldots,Y_{in_i}|X_{i1},\ldots,X_{in_i},\sum_{j=1}^{n_i}Y_{ij}=2\right)&\\ =\frac{\mathrm{pr}(Y_{i1},\ldots,Y_{in_i},\sum_{j=1}^{n_i}Y_{ij}=2|X_{i1},\ldots,X_{in_i})}{\mathrm{pr}(\sum_{j=1}^{n_i}Y_{ij}=2|X_{i1},\ldots,X_{in_i})}&\\ =\frac{\exp\{\beta(X_{i1}+X_{i2})\}\int\exp(2a_i+g_{i1}+g_{i2})d_i(\beta)\,dF(a,g)}{\sum_{k,l\in R_i}\exp\{\beta(X_{ik}+X_{il})\}\int\exp(2a_i+g_{ik}+g_{il})d_i(\beta)\,dF(a,g)}.&\quad(2{\cdot}4)\end{aligned}$$

The expressions in the numerator and the denominator of the last line are obtained from (2·2), and the summation is over all $n_i(n_i-1)/2$ pairs in the set R_i that consists of selections of two possible 'cases' from any of the n_i family members. The conditional likelihood function for m families is the product

$$L(Y_1,\ldots,Y_m,\theta)=\prod_{i=1}^{m}\frac{\exp\{\beta(X_{i1}+X_{i2})\}\int\exp(2a_i+g_{i1}+g_{i2})d_i(\beta)\,dF(a,g)}{\sum_{k,l\in R_i}\exp\{\beta(X_{ik}+X_{il})\}\int\exp(2a_i+g_{ik}+g_{il})d_i(\beta)\,dF(a,g)},\quad(2{\cdot}5)$$

where $\theta=(\mu,\beta,\lambda)$ and $Y_i=(Y_{i1},\ldots,Y_{in_i})$. Note that, in the absence of genetic effects, in which case $dF(a,g)$ assigns all mass to $g_i=(0,\ldots,0)$, $\sum_j Y_{ij}$ is the sufficient statistic for the family-specific intercept $\mu+a_i$, and the likelihood (2·5) reduces to

$$L(Y_1,\ldots,Y_m,\beta)=\prod_{i=1}^{m}\frac{\exp\{\beta(X_{i1}+X_{i2})\}}{\sum_{k,l\in R_i}\exp\{\beta(X_{ik}+X_{il})\}},\quad(2{\cdot}6)$$

just as in the case of conditional logistic regression (Cox, 1970, p. 45).

3. Results

3·1. *Bias calculations*

To estimate β in (2·1) from families with two affected members, we need to maximise the conditional likelihood given in (2·5). This is numerically challenging; see § 4. Standard conditional logistic regression based on (2·6) is computationally stable and rapid. We are thus interested in the properties of the estimator obtained from maximising (2·6), when the true model is the one including the random effects a and g.

Following an approach by Neuhaus et al. (1992), we show the following result.

Theorem 1. *If* $\beta=0$ *in* (2·5), *then the maximum likelihood estimators* $\hat{\beta}^*$ *based on the likelihood* (2·5), *but with a misspecified random effects distribution*, $G(a,g)$, *consistently estimate zero.*

Proof. Akaike (1973) and White (1982) show that the maximum likelihood estimator under the false model converges to the value $\theta^*=(\mu^*,\beta^*,\lambda^*)$ which minimises the Kullback–Leibler divergence between the true model F and the misspecified model G:

$$\theta^*=\arg\min_{\gamma}E_{X|\Sigma Y}E_{Y|X,\Sigma Y}\log\frac{\mathrm{pr}_F(Y|\theta,X,\sum Y)}{\mathrm{pr}_G(Y|\gamma,X,\sum Y)},\quad(3{\cdot}1)$$

where the expectation is taken with respect to the true model F. After differentiating the

above expression with respect to γ and some simplification, we see that θ^* has to satisfy

$$E_X\left\{\mathrm{pr}_F\left(\sum Y|\theta, X\right)\sum_y \mathrm{pr}_F\left(y|\theta, X, \sum Y\right)\frac{(d/d\gamma)\,\mathrm{pr}_G(y|\gamma, X, \sum Y)}{\mathrm{pr}_G(y|\gamma, X, \sum Y)}\right\}=0. \qquad (3{\cdot}2)$$

When $\beta=0$, the distribution of Y under the true model F does not depend on X, and equation (3·3) reduces to

$$\mathrm{pr}_F\left(\sum Y\right)\sum_y \mathrm{pr}_F\left(y|\sum Y\right)E_X\left\{\frac{(d/d\gamma)\,\mathrm{pr}_G(y|\gamma, X, \sum Y)}{\mathrm{pr}_G(y|\gamma, X, \sum Y)}\right\}=0.$$

After calculating the derivative with respect to the component of γ that corresponds to the coefficient of X and taking expectations, see the Appendix for details, it can be seen that the above score equation corresponding to β^* is satisfied for $\beta^*=0$ for all (μ, λ) and (μ^*, λ^*). The remaining score equations determine (μ^*, λ^*) in terms of (μ, λ). When $\beta\neq 0$, misspecification of the random effects distribution results in inconsistent estimators $\hat{\beta}^*$. We cannot solve equation (3·2) for β^* explicitly when $\beta\neq 0$ for a general G, because the calculations involve intractable expectations of nonlinear functions of X. □

From the Theorem we can derive the following corollary for conditional logistic regression, that is based on a specific choice of misspecified random effects distribution G, namely a distribution that assigns all mass to $g=(0, 0, \ldots, 0)$. This misspecification corresponds to ignoring the g_{ij} and using conditional logistic regression to estimate β.

COROLLARY 1. (i) *If* $\beta=0$ *in* (2·5), *then ignoring the random effects* g_{ij} *and using the standard conditional logistic likelihood* (2·6) *will give consistent estimators of* $\beta=0$.

(ii) *For small* $|\beta|\neq 0$, *the estimator* β^* *obtained from maximising* (2·6) *is biased toward the null, that is* $|\beta^*|<|\beta|$, *and the asymptotic relative bias is given by* (A·3) *in the Appendix.*

The proof of the Corollary is in the Appendix.

We note the related result that the score test of the hypotheses $H_0: \beta=0$ derived from the standard conditional regression has the correct size. The score test based on (2·6) is $\sum_i(X_{i1}+X_{i2}-E_i)/\{\sum V_i\}^{\frac{1}{2}}$, where $E_i=2\sum_{l,k}(X_{il}+X_{ik})/\{n_i(n_i-1)\}$ and $V_i=2/\{n_i(n_i-1)\}\{\sum_{k,l}(X_{ik}+X_{il})^2\}-E_i^2$ is the corresponding variance. Under the assumptions that $\beta=0$ and that the covariates are independent of a_i and g, $(X_{i1}, \ldots, X_{in_i})$ are exchangeable. Therefore, from standard theory for U-statistics, the distribution of the score statistic will tend to a standard normal distribution as the number of families increases, or, for a fixed number of families, as the n_i increase. Even though the standard score test based on (2·6) has proper size, it is theoretically less powerful than a score test based on (2·5).

3·2. *An approximation to the full likelihood for a rare disease*

Many problems that require an ascertainment correction concern a rare disease. In this setting μ is a large negative number. If the familial effect a, the environmental effect βX and the genetic effects g are small compared to $|\mu|$, then $\exp(\mu+a_i+g_{ij}+\beta X_{ij})\simeq 0$ and the disease probability given by (2·1) is small. Under these assumptions $d_i(\beta)\simeq 1$ and equation (2·4) is approximately

$$\frac{\exp\{\beta(X_{i1}+X_{i2})\}\int\exp(2a_i+g_{i1}+g_{i2})\,dF(a_i, g_i)}{\sum_{k,l\in R_i}\exp\{\beta(X_{ik}+X_{il})\}\int\exp(2a_i+g_{ik}+g_{il})\,dF(a_i, g_i)}.$$

The approximate conditional likelihood thus reduces to a weighted conditional logistic likelihood:

$$\mathrm{pr}\left(Y_{i1}, \ldots, Y_{in_i} | X_{i1}, \ldots, X_{in_i}, \sum_j Y_{ij} = 2 \right) = \frac{\exp\{\beta(X_{i1} + X_{i2})\} w_{12}}{\sum_{j,l \in R_i} \exp\{\beta(X_{ij} + X_{il})\} w_{jl}}, \qquad (3{\cdot}3)$$

where the weights $w_{kl} = \int \exp(2a_i + g_{ik} + g_{il})\, dF(a_i, g_i)$ do not depend on X. We call the solution to (3·3) a weighted conditional logistic regression estimator. If a_i and the g_{ij}'s are independent with joint distribution $F_a(a)F_g(g)$, the weights reduce to $w_{kl} = \int \exp(g_{ik} + g_{il})\, dF_g(g)$ and can be calculated from the moment generating function of g_{ik} and g_{il}. In this case, the likelihood (3·3) depends only on pairwise distribution functions of g_{ik} and g_{il} for all $l \neq k$. If all the family members have exchangeable genetic effects, for example if a family consists only of first-degree relatives, the weights in the numerator and denominator are all equal and cancel from the likelihood. In this case the likelihood reduces to conditional logistic regression. For a general family structure with nonexchangeable g_{ij}'s, as would arise in a pedigree with varying degrees of kinship, we obtain the following result.

Corollary 2. *If the rare disease approximation holds, the conditional logistic regression estimators obtained by maximising* (2·6) *are consistent for* β.

This result holds for arbitrary $F(a, g)$. The proof is in the Appendix.

3·3. *Explicit bias calculations for the sibling case-control design*

A design popular in association studies matches each case patient to an unaffected sibling. We calculate the bias explicitly for any value of β for the special case of a Bernoulli covariate $X \in \{0, 1\}$. Here the likelihood for m families is given by

$$\mathrm{pr}(Y_1, \ldots, Y_m, \theta) = \prod_{i=1}^{m} \mathrm{pr}(Y_{i1} | Y_{i1} + Y_{i2} = 1),$$

where the families are ordered such that Y_{i1} denotes the case, and $\theta = (\mu, \beta, \lambda)$.

Corollary 3. *If X is Bernoulli with* $\mathrm{pr}(X = 1) = p$, *then the estimator* $\hat{\beta}^*$ *obtained from maximising conditional logistic regression,*

$$\hat{\beta}^* = \arg\max_{\gamma} \prod_i \frac{\exp(\gamma X_{i1})}{\exp(\gamma X_{i1}) + \exp(\gamma X_{i2})},$$

converges to

$$\beta^* = \beta + \log \frac{w_1(1, 0) + w_2(0, 1)}{w_1(0, 1) + w_2(1, 0)}, \qquad (3{\cdot}4)$$

where

$$w_i(X_1, X_2) = \int \frac{\exp(\mu + a + g_i)\, dF(a, g)}{\{1 + \exp(\mu + a + g_1 + \beta X_1)\}\{1 + \exp(\mu + a + g_2 + \beta X_2)\}} \quad (i = 1, 2).$$

Moreover, $|\beta^*| \leqslant |\beta|$.

The proof is in the Appendix.

For a and g normally distributed and independent with respective variances σ_a^2 and

σ_g^2 and with $\text{corr}(g_1, g_2) = 0{\cdot}5$, we plot $\beta - \beta^*$ against σ_g^2 for $\beta = 1$ in Fig. 1. Different curves correspond to different values of μ and σ_a^2. The bias increases with the value of σ_g^2 and is larger for $\mu = -2$ than for $\mu = -5$. For $\mu = -5$ and $\sigma_g^2 = 3$, $\beta^* - \beta = -0{\cdot}105$ for $\sigma_a^2 = 0$, and $-0{\cdot}156$ for $\sigma_a^2 = 3$. For $\mu = -2$ and $\sigma_g^2 = 3$, $\beta^* - \beta = -0{\cdot}194$ for $\sigma_a^2 = 0$, and $-0{\cdot}197$ for $\sigma_a^2 = 3$.

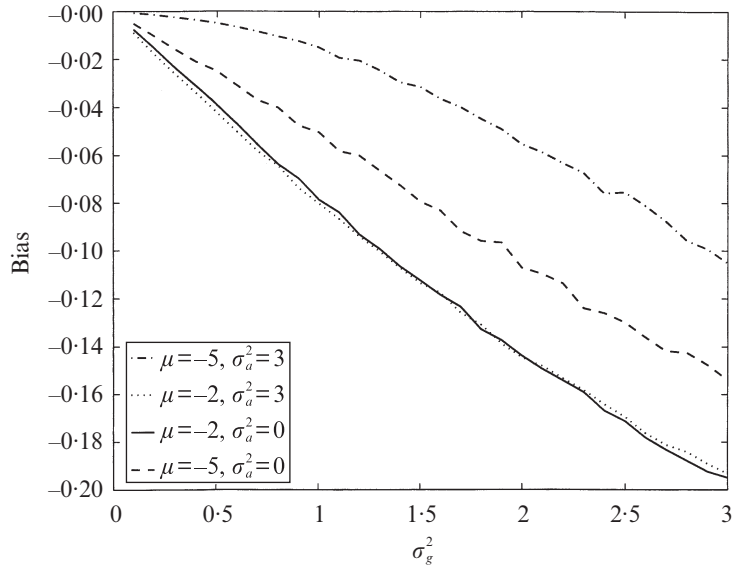

Fig. 1. Plot of bias, $\beta^* - \beta$, versus σ_g^2 for various values of μ and σ_a^2.

4. Simulation study

4·1. *Preamble*

We use simulated data to study the behaviour of the estimators based on the random effects model (2·5), conditional logistic regression (2·6) and weighted conditional logistic regression (3·3), and compare the results to the bias approximation given in (A·3). The simulated datasets represent 100 families. For simplicity we assume that all families have the same size, $n_i = 6$, and the same pedigree structure, described in § 4·2.

4·2. *Genetic random effects model*

Fisher (1918) showed that the genetic variance for a trait can be divided into two components, the additive genetic variance, which results from differences between homozygotes, and the dominance variance, which results from specific effects of various alleles in heterozygotes. We assume that the genetic variance for the latent liabilities g can be similarly divided. Following Fisher, and assuming no dominance component of the variance, we model the covariance matrix of the g's in the ith family, Σ_i, as a function of the degree of kinship between members in the family:

$$\text{cov}(g_{ij}, g_{il}) = (\Sigma_i)_{j,l} = \frac{\sigma_g^2}{2^{k(j,l)}}. \tag{4·1}$$

Here $k(j, l)$ denotes the degree of kinship between members j and l in the ith family. For

example, $k(j, j) = 0$, and $k(j, l) = 1$ if j and l are first-degree relatives, e.g. siblings, and $k(j, l) = 2$ if j and l represent a grandparent and a grandchild, or an aunt and a nephew. Thus for each extra generation the covariance is multiplied by a factor of $\frac{1}{2}$. For unrelated members of the family, such as spouses, $k(j, l) = \infty$ and $(\Sigma_i)_{j,l} = 0$. The covariance matrix corresponding to the family structure used in the example presented in Table 1, below, is given by

$$\Sigma = \sigma_g^2 \begin{pmatrix} 1{\cdot}0 & 0{\cdot}0 & 0{\cdot}5 & 0{\cdot}5 & 0{\cdot}25 & 0{\cdot}25 \\ 0{\cdot}0 & 1{\cdot}0 & 0{\cdot}5 & 0{\cdot}5 & 0{\cdot}25 & 0{\cdot}25 \\ 0{\cdot}5 & 0{\cdot}5 & 1{\cdot}0 & 0{\cdot}5 & 0{\cdot}25 & 0{\cdot}25 \\ 0{\cdot}5 & 0{\cdot}5 & 0{\cdot}5 & 1{\cdot}0 & 0{\cdot}5 & 0{\cdot}5 \\ 0{\cdot}25 & 0{\cdot}25 & 0{\cdot}25 & 0{\cdot}5 & 1{\cdot}0 & 0{\cdot}5 \\ 0{\cdot}25 & 0{\cdot}25 & 0{\cdot}25 & 0{\cdot}5 & 0{\cdot}5 & 1{\cdot}0 \end{pmatrix}. \tag{4·2}$$

This covariance matrix corresponds to the following ordering of family members: mother, father, offspring 1, offspring 2, child 1 of offspring 2 and child 2 of offspring 2. We call this pedigree structure 1. For the simulations of Table 2, below, we assumed that each family consisted of mother, father and four children, representing pedigree structure 2, that is $(\Sigma_i)_{j,l} = 0$ for $j = 1$ and $l = 2$, and $(\Sigma_i)_{j,l} = \sigma_g^2/2$ for all other $j \neq l$. Although we do not use it in this paper, the covariance matrix of a genetic random effects model that allows for an additive and a dominant component is $(\Sigma_i)_{j,j} = \sigma_g^2 + \sigma_d^2$, $(\Sigma_i)_{j,l} = \frac{1}{2}\sigma_g^2 + \frac{1}{4}\sigma_d^2$ if j and l are siblings and $(\Sigma_i)_{j,l} = \sigma_g^2/2^{k(j,l)}$ for all other types of relative.

In what follows we will assume that a_i and the g_{ij}'s are normally distributed and a_i is independent of the g_{ij}'s. Normality of g_{ij} is a reasonable assumption if the genetic liability is influenced by many genes (Fisher, 1918). Note that, if the rare disease assumption holds in this setting, the weighted conditional logistic regression estimators $\hat{\beta}$ and $\hat{\sigma}_g^2$ obtained from (3·3) are asymptotically uncorrelated and independent, which can be seen by calculating the off-diagonal term in the Fisher information matrix, namely

$$E\left[\frac{\partial}{\partial\beta}\frac{\partial}{\partial\sigma_g^2}\log\frac{w_{12}\exp\{\beta(X_1+X_2)\}}{\sum w_{ij}\exp\{\beta(X_i+X_j)\}}\right] = 0,$$

where $w_{ij} = \exp(\sigma_g^2/2^{k(i,j)})$.

4·3. *Numerical methods for the random effects model and weighted conditional logistic regression*

The parameters in the full random effects likelihood were estimated by direct maximisation. To evaluate the integrals in the conditional likelihood function (2·5), we used Monte Carlo integration. For each family, we drew independent, identically distributed samples a_i and $(g_{i1}, \ldots, g_{i6})$, for $i = 1, \ldots, N$, and used the approximation

$$\int \frac{\exp(2a + g_1 + g_2)}{\prod_{j=1}^{6}\{1 + \exp(\mu + a + \beta X_j + g_j)\}}\, dF(a, g) \doteq \frac{1}{N}\sum_{k}^{N} \frac{\exp(2a_k + g_{k1} + g_{k2})}{\prod_{j=1}^{6}\{1 + \exp(\mu + a_k + \beta X_j + g_{kj})\}}.$$

We chose $N = 100$ and used the same Monte Carlo sample for the numerator and denominator of the conditional likelihood of each family to ensure that the conditional likelihood was smooth in β. Different families were evaluated using independent Monte Carlo samples. The advantage of Monte Carlo integration over Gaussian quadrature is that the

required computations increase only linearly with the dimension of the integral, while the numerical effort for Gaussian quadrature increases exponentially with the dimension of the integral.

It is not possible to estimate μ and σ_a^2 from these data if $\sigma_g^2 = 0$, see equations (2·5) and (2·6), or if the disease is rare; see equation (3·3) and recall that w_{jl} depends only on σ_g^2 under independence of a_i from g_{ij}. Even with $\sigma_g^2 > 0$, the ascertainment scheme yields little information on μ and σ_a^2, and the profile likelihood in σ_a^2 is usually very flat in our numerical studies. To overcome numerical instability resulting from lack of information on μ and σ_a^2, and because we are primarily interested in inference on β, we replaced the four-dimensional maximisation problem by a double-grid search and a two-dimensional maximisation. For each $\mu \in \{-7, -6{\cdot}5, -6, \ldots, -3{\cdot}5, -3, -2{\cdot}5, -2\}$ and every $\sigma_a^2 \in \{0, 0{\cdot}5, 1, 1{\cdot}5, 2, \ldots, 19, 19{\cdot}5, 20\}$ we maximised the likelihood as a function of β and σ_g^2. The final estimator was the (β, σ_g^2) pair that yielded the biggest likelihood over the whole grid. Confidence intervals for β were based on the likelihood ratio statistic from the full model (2·5). To see if such a confidence interval covers β, one only needs to evaluate the likelihood at $\hat{\beta}$ and at β.

The maximisation of the likelihood (3·3) corresponding to weighted conditional logistic regression is straightforward in the normal setting, as the weights can be calculated explicitly, as $w_{jl} = \exp(\sigma_g^2/2^{k(j,l)})$. Thus, standard optimisation programs were used to maximise (3·3) jointly over β and σ_g^2.

4·4. *Results*

Table 1 presents simulation results for $\beta = 1$, $\mu = -5$ or -2, a Bernoulli covariate $X \in \{0, 1\}$ with $p = 0{\cdot}5$, and various choices for values of σ_a^2 and σ_g^2 using the covariance matrix (4·2). The random effects model (2·5) yielded nearly unbiased estimates for $\beta = 1$ and near nominal 95% coverage of β for likelihood ratio based confidence intervals for each of the parameter combinations studied in Table 1. Estimates of σ_g^2 had much larger coefficients of variation than estimates of β, and there was a tendency to overestimate σ_g^2 when the true value was 0·5. Unreported confidence intervals based on the Wald statistic for β and σ_g^2 had slightly subnominal coverage.

In the cases where the rare-disease approximation holds, namely $\mu = -5$ with $\sigma_a^2 = 0{\cdot}5$ or 1·0 and $\sigma_g^2 = 0{\cdot}5$ or 1·0, the estimates of β from conditional logistic regression and weighted conditional logistic regression were nearly unbiased and the corresponding Wald confidence intervals had nearly nominal coverage; see Table 1. Otherwise, the bias is noticeable and in good agreement with the analytical approximation (A·3); see Table 3. For example, for $\mu = -5$, $\sigma_a^2 = 3$ and $\sigma_g^2 = 5$, the bias of conditional logistic regression was $-0{\cdot}286$, whereas (A·3) yielded $-0{\cdot}285$. The corresponding confidence intervals for β had coverage 0·579 and 0·657 for weighted conditional logistic regression and conditional logistic regression respectively. For $\mu = -5$, $\sigma_a^2 = 1$ and $\sigma_g^2 = 5$, the empirical bias of conditional logistic regression was $-0{\cdot}237$, compared to a value from (A·3) of $-0{\cdot}265$.

As the values of σ_a^2 and σ_g^2 declined, relative to $|\mu|$, the performance of weighted conditional logistic regression and conditional logistic regression improved. For example, with $\mu = -5$, $\sigma_a^2 = 1$ and $\sigma_g^2 = 1$ the empirical bias of conditional logistic regression was $-0{\cdot}057$, compared to $-0{\cdot}045$ from the analytical approximation (A·3). Weighted conditional logistic regression and conditional logistic regression exhibit negligible bias and near nominal coverage for β with $\mu = -5$, $\sigma_a^2 = 0{\cdot}5$ and $\sigma_g^2 = 0{\cdot}5$.

For more common diseases, with $\mu = -2$, the biases in weighted conditional logistic

Table 1. *Simulation results for estimation of β and σ_g^2, with estimated standard errors in parentheses, for pedigree structure* 1

	Two-level random effects model			WCLR (upper entries) CLR (lower entries)		
$\mu, \sigma_a^2, \sigma_g^2$	Mean $\hat{\beta}$	Coverage	Mean $\hat{\sigma}_g^2$	Mean $\hat{\beta}$	Coverage	Mean $\hat{\sigma}_g^2$
−5, 3·0, 5·0	0·957 (0·330)	0·947	4·848 (4·248)	0·692 (0·181) 0·714 (0·178)	0·579 0·675	1·626 (0·793)
−5, 1·0, 5·0	0·975 (0·302)	0·940	3·549 (4·199)	0·765 (0·189) 0·763 (0·186)	0·710 0·750	1·603 (0·682)
−5, 1·0, 1·0	1·038 (0·258)	0·930	1·257 (1·437)	0·944 (0·193) 0·943 (0·192)	0·930 0·940	0·759 (0·613)
−5, 0·5, 0·5	1·099 (0·217)	0·927	0·993 (0·970)	1·020 (0·193) 1·019 (0·192)	0·963 0·973	0·539 (0·567)
−5, 0·0, 0·0	1·024 (0·197)	0·941	0·380 (0·707)	0·999 (0·190) 0·998 (0·192)	0·948 0·947	0·259 (0·405)
−2, 3·0, 5·0	1·000 (0·344)	0·927	4·373 (4·851)	0·693 (0·177) 0·691 (0·173)	0·627 0·627	0·962 (0·580)
−2, 1·0, 5·0	1·007 (0·321)	0·940	5·133 (4·237)	0·707 (0·186) 0·705 (0·185)	0·640 0·640	1·106 (0·700)
−2, 1·0, 1·0	1·040 (0·223)	0·950	1·401 (1·249)	0·918 (0·151) 0·917 (0·151)	0·920 0·920	0·465 (0·520)
−2, 0·5, 0·5	1·067 (0·291)	0·930	1·229 (2·039)	0·961 (0·207) 0·960 (0·206)	0·920 0·920	0·387 (0·534)
−2, 0·0, 0·0	1·035 (0·251)	0·950	0·519 (1·584)	0·998 (0·184) 0·997 (0·185)	0·950 0·950	0·251 (0·360)

WCLR, weighted conditional logistic regression; CLR, conditional logistic regression; Coverage, coverage for β.

Table 2. *Simulation results for estimation of β and σ_g^2, with estimated standard errors in parentheses, for pedigree structure* 2

	Two-level random effects model			WCLR (upper entries) CLR (lower entries)		
$\mu, \sigma_a^2, \sigma_g^2$	Mean $\hat{\beta}$	Coverage	Mean $\hat{\sigma}_g^2$	Mean $\hat{\beta}$	Coverage	Mean $\hat{\sigma}_g^2$
−5, 3·0, 5·0	0·960 (0·296)	0·916	3·728 (3·527)	0·756 (0·177) 0·754 (0·176)	0·720 0·738	4·044 (6·775)
−5, 1·0, 1·0	1·025 (0·246)	0·959	1·693 (2·389)	0·925 (0·183) 0·924 (0·183)	0·928 0·940	1·365 (3·007)
−5, 0·5, 0·5	1·065 (0·212)	0·934	1·059 (1·210)	0·998 (0·180) 0·998 (0·179)	0·962 0·962	0·697 (0·805)
−2, 3·0, 5·0	0·981 (0·327)	0·933	4·937 (3·611)	0·706 (0·186) 0·705 (0·185)	0·587 0·615	3·006 (5·957)
−2, 1·0, 1·0	1·043 (0·257)	0·930	1·738 (1·831)	0·906 (0·192) 0·905 (0·191)	0·887 0·890	0·759 (1·786)
−2, 0·5, 0·5	1·046 (0·234)	0·973	1·456 (1·556)	0·929 (0·172) 0·929 (0·171)	0·927 0·927	0·814 (2·801)

WCLR, weighted conditional logistic regression; CLR, conditional logistic regression; Coverage, coverage for β.

Table 3. *Relative bias* $(\hat{\beta}-\beta)/\beta$ *computed from* (A·3) *and corresponding average estimated relative bias from simulations*

Pedigree structure 1			Pedigree structure 2		
$\mu, \sigma_a^2, \sigma_g^2$	Bias from (A·3)	Bias from simulations	$\mu, \sigma_a^2, \sigma_g^2$	Bias from (A·3)	Bias from simulations
−5, 3·0, 5·0	−0·2850	−0·2861	−5, 3·0, 5·0	−0·2559	−0·2440
−5, 1·0, 5·0	−0·2651	−0·2372	−5, 1·0, 1·0	−0·0406	−0·0750
−5, 1·0, 1·0	−0·0458	−0·0576	−5, 0·5, 0·5	−0·0082	−0·0070
−5, 0·5, 0·5	−0·0110	−0·0102	−2, 3·0, 5·0	−0·2839	−0·2940
−5, 0·0, 0·0	0·0000	−0·0020	−2, 1·0, 1·0	−0·0786	−0·0940
−2, 3·0, 5·0	−0·3159	−0·3090	−2, 0·5, 0·5	−0·0401	−0·0510
−2, 1·0, 5·0	−0·3150	−0·2959			
−2, 1·0, 1·0	−0·0935	−0·0837			
−2, 0·5, 0·5	−0·0464	−0·0402			
−2, 0·0, 0·0	0·0000	−0·0031			

regression and conditional logistic regression were more pronounced. Even for $\sigma_a^2=\sigma_g^2=1$, there remained an estimated bias for β of $-0{\cdot}083$ for conditional logistic regression, and for $\sigma_a^2=\sigma_g^2=0{\cdot}5$ an estimated bias of $-0{\cdot}04$. These biases were again in good agreement with the respective biases calculated from (A·3), namely $-0{\cdot}09$ and $-0{\cdot}04$. The weighted conditional logistic regression procedure underestimated σ_g^2 by a factor of 2 or more for each of the sets of parameter values in Table 1.

The random effects model (2·5) yielded reliable inference on β for pedigree structure 2, see Table 2, just as for pedigree structure 1, see Table 1, and the tendency to overestimate $\sigma_g^2=0{\cdot}5$ persisted with pedigree structure 2. The magnitudes of the bias in β for weighted conditional logistic regression and conditional logistic regression with pedigree structure 2 were similar to those with pedigree structure 1. Estimates of σ_g^2 from weighted conditional logistic regression exhibited less downward bias with pedigree structure 2 than with pedigree structure 1, however. For example, for $\mu=-5$, $\sigma_a^2=3$ and $\sigma_g^2=5$, the average estimate of σ_g^2 was 4·04 for pedigree structure 2, see Table 2, compared to only 1·63 for pedigree structure 1, see Table 1. Likewise, for $\mu=-2$, $\sigma_a^2=3$ and $\sigma_g^2=5$, the mean estimate of σ_g^2 from weighted conditional logistic regression was 3·00 for pedigree structure 2 compared to 0·96 for pedigree structure 1.

The performance of the estimates of σ_g^2 based on the random effects model (2·5) deteriorated as σ_g^2 got smaller. This is because the estimator is constrained to $\sigma_g^2\geqslant 0$. For example, for $\mu=-5$, $\sigma_a^2=0{\cdot}5$ and $\sigma_g^2=0{\cdot}5$ in Table 1, the average estimate of σ_g^2 was 0·99, which corresponds to a relative bias of 99%. For $\mu=-2$, $\sigma_a^2=0{\cdot}5$ and $\sigma_g^2=0{\cdot}5$ the relative bias was even more pronounced, as the average estimate of σ_g^2 was 1·23. For $\mu=-5$, $\sigma_a^2=0$ and $\sigma_g^2=0$, the average estimate of σ_g^2 was 0·380, but the average estimate of β, 1·024, indicated that $\hat{\beta}$ is nearly unbiased.

5. A REAL EXAMPLE

The data derived from 65 families with $n_i=6$ members, and each family had exactly two members affected with nasopharyngeal carcinoma. These data represent a subset of the families that participated in the complete study. The covariates we considered were $X_1=1$ for 'ever smoker' and 0 for 'never smoker', $X_2=1$ for female and 0 for male, and two age-group indicators, $X_3=1$ for age $\leqslant 46$ years, $X_3=0$ otherwise, and $X_4=1$ for age 46–57 years, $X_4=0$ otherwise. The $\geqslant 57$ age group is the reference group. The conditional

Table 4. *Nasopharyngeal carcinoma data results for estimation of β and σ_g^2, with estimated standard errors in parentheses*

Parameter	Two-level random effects model	WCLR	CLR
β_1	0·1536 (0·2684)	0·1643 (0·2603)	0·1706 (0·2606)
β_2	−0·8794 (0·2581)	−0·8597 (0·2517)	−0·8582 (0·2525)
β_3	−0·2666 (0·3123)	−0·2742 (0·3046)	−0·2736 (0·3053)
β_4	0·5984 (0·2986)	0·5895 (0·2871)	0·5872 (0·2869)
σ_g^2	0·6511 (0·7511)	0·3544 (0·8502)	Not available
Loglikelihood	−162·965	−163·455	−163·544

WCLR, weighted conditional logistic regression; CLR, conditional logistic regression.

likelihood model, weighted conditional logistic regression and conditional logistic regression were fitted to the data.

To fit model (2·5), we chose the random genetic effects distribution to be multivariate normal with an additive covariance structure, specified in (4·1). Different covariance matrices were required for different family structures. The random familial effect a_i was also assumed to be normally distributed and independent of the g_{ij}. The numerical methods in §§ 4·2 and 4·3 were used to maximise the loglikelihood for model (2·5). The grid search on μ and σ_a^2 resulted in the estimates $\hat{\mu} = -3{\cdot}5$ and $\hat{\sigma}_a^2 = 0$.

Table 4 presents the point estimates of β and standards errors, in parentheses, for each model. All three models suggest that two of the covariates have statistically significant nonnull effects, namely gender, X_2, and the 46–57 age group, X_4.

The weighted conditional logistic regression and conditional logistic regression estimates of the coefficients for these variables are slightly smaller in magnitude than the estimates based on (2·5), as one would anticipate from Corollary 1. The estimates of β_2 and β_4 are −0·8794 and 0·5984 respectively, for the random effects model, and −0·8582 and 0·5872 respectively, for conditional logistic regression. The coefficients for X_1, smoking, and X_4, the 46–57 age group, are not significantly different from zero. The weighted conditional logistic regression and conditional logistic regression estimates of β for these two covariates are not smaller in magnitude than estimates from (2·5). This may reflect random variation in the estimates or noise introduced by the Monte Carlo integration in a setting with small values of β and small genetic variation, σ_g^2.

A notable feature of Table 4 is the close agreement among the estimates of β for all the models. This may reflect the fact that nasopharyngeal carcinoma is a relatively rare disease in Taiwan, and that σ_g^2 is comparatively small. In this situation the use of the much simpler conditional logistic regression or weighted conditional logistic regression procedure may be justified with little danger of serious bias. The loglikelihood of (2·5), −162·965, only decreased to −163·544 for conditional logistic regression and −163·455 for weighted conditional logistic regression, which does not suggest a need for the more complex random effects model with three more parameters than conditional logistic regression.

Before drawing firm conclusions on which model to use, however, the full dataset needs to be examined with a dominant as well as an additive component to the genetic covariance, and one needs to study covariates with larger log-relative-odds, β, for which the bias towards the null may be more striking.

Our preliminary findings are consistent with earlier work demonstrating lower risk in women and elevated risk in the 46–57 year age group in Taiwan (Hildesheim & Levine, 1993). Cheng et al. (1999) also found a nonsignificant increase in risk associated with smoking classified as ever or never smoker.

6. Discussion

We have proposed methods based on a two-stage random effects model (2·1) to account for genetic effects in family studies. The conditional likelihood (2·4) takes into account the ascertainment scheme.

Our analytic work and simulation studies show that conditional logistic regression can yield biased estimators of β. There is no bias, however, when $\beta = 0$, and the bias will be small when the genetic random effects are small, or when $\mu + a_i + g_{ij} + \beta X_{ij} \ll 0$ for all subjects. If $g_{ij} = 0$ for all i and j, then model (2·1) reduces to a logistic model with a random intercept, for which conditional logistic regression yields unbiased results of β. Gail et al. (1984) and Neuhaus (1993) showed how unconditional logistic regression assuming a constant intercept leads to biased estimators of β when $g_{ij} = 0$. The condition $\mu + a_i + g_{ij} + \beta X_{ij} \ll 0$ is satisfied if the disease is rare and βX_{ij}, a_i and g_{ij} are small compared to $|\mu|$. In this setting, and under the normal random effects model in § 4·1, the conditional likelihood (2·4), which is well approximated by (3·3), provides little information on μ and σ_a^2. If, in addition, the bivariate distributions $F(g_{il}, g_{ik})$ are all equal, as might occur in a study of siblings, then the conditional likelihood (2·4) provides little information on the parameters of F_g; see equation (3·3). In each of these cases, however, conditional logistic regression can be relied on to provide valid inference on β, the parameter of primary interest in this paper. Otherwise, one is forced to use more complex methods based on the conditional random effects likelihood (2·4). This approach has its own drawbacks. The computations are difficult and specialised. Although the Monte Carlo approach worked well in our examples, larger Monte Carlo samples or other methods may be needed for larger pedigrees. More fundamentally, often one does not know the precise nature of the genetic influences and hence the distribution of g_{ij}. One approach might be to compare results from conditional logistic regression with the results from model (2·5), using the distribution for the g_{ij}'s in § 4. If the results are similar, one can probably rely on conditional logistic regression. If not, a broader range of models for the distribution of the g_{ij} needs to be explored to make sure that inference on β from (2·5) is robust.

If families were selected at random, then standard Gibbs sampling methods could be used to estimate the parameters of model (2·1). Since our ascertained families do not represent the entire space of possible outcomes, however, standard application of the Gibbs sampler yields biased estimates of the parameters in model (2·1); see Burton et al. (2000).

If the sole purpose of a study were to investigate the effects of measured covariates on disease risk, then other designs such as cohort or case-control designs could be used. Analyses would usually be based on the fixed effects 'population averaged' model described in § 2, with a_i and g_{ij} omitted. If the random effects model (2·1) is correct, however, such fixed effects models would result in biased estimators of the covariate effects at the individual level (Zeger et al., 1988).

Two-stage nested random effects models for dichotomous outcomes also arise in other settings. For example, model (2·1) could be used in longitudinal studies, where Y_{ij} denotes the jth repeat of a measurement on the ith individual. Another application would be to matched case-control studies in which the case and control have nonidentical unmeasured exposures to factors other than the measured exposure X.

Acknowledgement

We thank Allan Hildesheim, for bringing the problem to our attention, Ray Carroll, Lynn Goldin, Paul Burton, Alisa Goldstein, Louise Ryan and the referees, for many helpful

comments, and Edward Suh from the Center for Information Technology, National Institutes of Health, for help with the simulations.

Appendix

Proofs

Calculation of the derivative for proof of Theorem 1. Let γ denote the coefficient of X. We now show that, if $\beta = 0$, then

$$E_X\{d/d\gamma \operatorname{pr}_G(y|\gamma, X, \sum Y)/\operatorname{pr}_G(y|\gamma, X, \sum Y)\} = 0$$

is satisfied for $\gamma = 0$. Note that, if $\gamma = 0$, pr_G does not depend on X, and thus it suffices to show that $E_X\{d/d\gamma \operatorname{pr}_G(y|\gamma, X, \sum Y)\} = 0$ for $\gamma = 0$. Recall that $\operatorname{logit}\{p_{ij}(\gamma)\} = \mu + a_i + g_{ij} + \gamma X_{ij}$. Let

$$N(\gamma) = \sum_{k,l \in R_i} \exp\{\gamma(X_{ik} + X_{il})\} \int \exp(2a_i + g_{ik} + g_{il}) d_i(\gamma)\, dG(a, g),$$

with $d_i(\gamma)$ defined in (2·3). Then $d/d\gamma \operatorname{pr}_G(y|\gamma, X, \sum Y) = A(\gamma)/N(\gamma) - B(\gamma)/N^2(\gamma)$, where

$$A(\gamma) = \exp\{\gamma(X_{i1} + X_{i2})\} \int \exp(2a_i + g_{i1} + g_{i2}) d_i(\gamma) \left\{X_{i1} + X_{i2} - \sum_j X_{ij} p_{ij}(\gamma)\right\} dG(a, g),$$

$$B(\gamma) = \exp\{\gamma(X_{i1} + X_{i2})\} \int \exp(2a_i + g_{i1} + g_{i2}) d_i(\gamma)\, dG(a, g)$$

$$\times \sum_{k,l} \int \exp(2a_i + g_{ik} + g_{il}) d_i(\gamma) \exp\{\gamma(X_{ik} + X_{il})\} \left\{X_{ik} + X_{il} - \sum_j X_{ij} p_{ij}(\gamma)\right\} dG(a, g),$$

with p_{ij} defined in (2·1) as a function of γ. As $N(0)$ is independent of X and, since $E(X) = 0$, $E\{A(0)\} = 0$ and $E\{B(0)\} = 0$ the equation $E_X\{d/d\gamma \operatorname{pr}_G(y|\gamma, X, \sum Y)\} = 0$ is satisfied for $\gamma = 0$. Calculation of the second derivative shows that $\gamma = 0$ gives a minimum. Thus for $\beta = 0$ the left-hand side of equation (3·2) corresponding to the coefficient of X is zero independent of the choice of $G(a, g)$. For $\beta \neq 0$, equation (3·2) is typically not satisfied for $\gamma = \beta$ and thus yields inconsistent estimators of the true β.

Proof of Corollary 1 (ii). Without loss of generality let $E(X) = 0$. For brevity we omit the family index, set

$$w_{ij}(X) \equiv \int \exp(2a + g_i + g_j) d(\beta)\, dF(a, g)$$

and let $\operatorname{logit}\{p_m(\gamma)\} = \mu + a + g_m + \gamma X_m$. Note that

$$w'_{ij}(X) = \frac{d}{d\beta} w_{ij}(X) = -\sum_m X_m \int \exp(2a + g_i + g_j) d(\beta) p_m(\beta)\, dF(a, g).$$

Then β^* is the solution of equation (3·2), which can be written as

$$E_X \sum_{l,m \in R} (X_l + X_m) w_{lm}(X) \exp\{\beta(X_l + X_m)\}$$

$$= E_X \frac{\sum_{l,m \in R} w_{lm}(X) \exp\{\beta(X_{ln} + X_m)\} \sum_{i,j \in R} (X_i + X_j) \exp\{\beta^*(X_i + X_j)\}}{\sum_{k,l \in R} \exp\{\beta^*(X_k + X_l)\}}. \quad \text{(A·1)}$$

Linearising the right-hand side of equation (A·1) around the true value β we obtain

$$E_X \sum_{l,m} (X_l + X_m) w_{lm}(X) \exp\{\beta(X_l + X_m)\}$$
$$- E_X \left[\sum_{l,m} w_{lm}(X) \exp\{\beta(X_l + X_m)\} \frac{\sum_{l,m} (X_l + X_m) \exp\{\beta(X_l + X_m)\}}{\sum_{j,k} \exp\{\beta(X_j + X_k)\}} \right]$$
$$= (\beta^* - \beta) E_X \sum_{l,m} w_{lm}(X) \exp\{\beta(X_l + X_m)\}$$
$$\times \left(\frac{\sum_{l,m} (X_l + X_m)^2 \exp\{\beta(X_l + X_m)\}}{\sum_{j,k} \exp\{\beta(X_j + X_k)\}} - \left[\frac{\sum_{l,m} (X_l + X_m) \exp\{\beta(X_l + X_m)\}}{\sum_{j,k} \exp\{\beta(X_j + X_k)\}} \right]^2 \right). \quad \text{(A·2)}$$

With

$$q_{lm} \equiv \frac{\exp\{\beta(X_l + X_m)\}}{\sum_{j,k} \exp\{\beta(X_j + X_k)\}}, \quad Y \equiv \left[\sum_{l,m} w_{lm}(X) \exp\{\beta(X_l + X_m)\} \right]^{1/2},$$

the square root being well defined as the expression is always positive, we rewrite the coefficient of $(\beta^* - \beta)$ as $G(\beta) = E_X[\sum_{l,m} Y^2 (X_l + X_m)^2 q_{lm} - \{\sum_{l,m} Y(X_l + X_n) q_{lm}\}^2]$, which has the form of a variance, and thus is always positive. Denote the left-hand side of equation (A·2) by $H(\beta)$ and let $\beta > 0$. Since $E(X) = 0$, we have $H(0) = 0$. Straightforward calculation shows that

$$H'(0) = \frac{E(X^2)}{n} \sum_{l,m} \int \exp(2a + g_l + g_m) d(0) \left\{ -np_l(0) - np_m(0) + 2 \sum_{i=1}^{n} p_i(0) \right\} dF(a, g) < 0.$$

Thus, for small values of β, $H(\beta) < 0$, which implies that $(\beta^* - \beta)$ is negative. A similar argument with $\beta < 0$ shows that $(\beta^* - \beta)$ is positive, which proves the claim.

If we rewrite equation (A·2) in the form $\beta^* - \beta = H(\beta)/G(\beta) \doteq \beta H'(0)/G(0)$ and evaluate the expectations with respect to X, an approximation to the relative bias for small β is given by

$$\frac{\beta^* - \beta}{\beta} = \frac{\sum_{l,m} \int \exp(2a + g_l + g_m) d(0) \{ -np_l(0) - np_m(0) + 2 \sum_{i=1}^{n} p_i(0) \} \, dF(a, g)}{2(n-2) \sum_{l,m} \int \exp(2a + g_l + g_m) d(0) \, dF(a, g)}. \quad \text{(A·3)}$$

We note that the bias does not depend on $E(X^2)$, which cancels out when $E(X^2)$ divides $H'(0)$.

Based on results from Table 3 and on comments of a reviewer, we conjecture that the bias approximation is monotone in the parameters μ, σ_a^2 and σ_g^2, but we were unable to prove it.

Proof of Corollary 2. For brevity we omit the family index and set

$$w_{ij} \equiv \int \exp(2a + g_i + g_j) \, dF(a, g).$$

Noting that the true model is now given by (3·3), we find the estimator as the solution of equation (3·2), which can be rewritten as

$$E_X \left(\left[\sum_{n,m} (X_n + X_m) w_{nm} \exp\{\beta(X_n + X_m)\} \sum_{k,l} \exp\{\beta^*(X_k + X_l\} \right.\right.$$
$$\left.\left. - \sum_{n,m} w_{nm} \exp\{\beta(X_n + X_m)\} \sum_{i,j} (X_i + X_j) \exp\{\beta^*(X_i + X_j)\} \right] \Big/ \sum_{k,l} \exp\{\beta^*(X_k + X_l)\} \right) = 0.$$

Noting that the weights w_{ij} do not depend on X, we see that for $\beta^* = \beta$ the above equation is satisfied, because the exchangeability of the X's in the population implies that

$$E_X[(X_n + X_m) \exp\{\beta(X_n + X_m + X_k + X_l)\}] = E_X[(X_k + X_l) \exp\{\beta(X_n + X_m + X_k + X_l)\}].$$

Proof of Corollary 3. For this setting equation (3·1) reduces to

$$E_X\left[X_1 w_1(X_1, X_2)\exp(\beta X_1) + X_2 w_2(X_1, X_2)\exp(X_2) - \frac{\{w_1(X_1, X_2)\exp(\beta X_1) + w_2(X_1, X_2)\exp(\beta X_2)\}\{X_1\exp(\beta^* X_1) + X_2\exp(\beta^* X_2)\}}{\exp(\beta^* X_1) + \exp(\beta^* X_2)}\right] = 0.$$

The definition of $w_i(X_1, X_2)$ is stated in Corollary 3. Evaluating the expectation results in (3·4). Next we determine the sign of the bias term. Straightforward calculation yields

$$\begin{aligned} w_1(1,0) + w_2(0,1) - w_1(0,1) - w_2(1,0) &= \int \exp(2a + 2\mu)\{1 - \exp(\beta)\}\{\exp(g_1) - \exp(g_2)\}^2 \\ &\quad \times [\{1 + \exp(\mu + a + g_1 + \beta)\}\{1 + \exp(\mu + a + g_2)\}\{1 + \exp(\mu + a + g_1)\} \\ &\quad \times \{1 + \exp(\mu + a + g_2 + \beta)\}]^{-1}\, dF(a, g). \end{aligned}$$

Thus $w_1(1,0) + w_2(0,1) - w_1(0,1) - w_2(1,0)$ is negative for $\beta > 0$ and positive for $\beta < 0$. As $w_1(0,1) + w_2(1,0) > 0$, we obtain that $\{w_1(1,0) + w_2(0,1)\}/\{w_1(0,1) + w_2(1,0)\} > 1$ for $\beta < 0$, and < 1 for $\beta > 0$, which corresponds to a bias towards the null in equation (3·4).

REFERENCES

AKAIKE, H. (1973). Information theory and extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Ed. B. N. Petrov and F. Czaki, pp. 267–91. Budapest: Akademiai Kiado.

AMOS, C. I. (1993). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54**, 535–43.

BURTON, P. R. PALMER, L. J., JACOBS, K., KEEN, K. J., OLSON, J. M. & ELSTON, R. C. (2000). Ascertainment adjustment: Where does it take us? *Am. J. Hum. Genet.* **76**, 1505–14.

COX, D. R. (1970). *Analysis of Binary Data*. London: Methuen.

CHENG, Y. J., HILDESHEIM, A., HSU, M. M., CHEN, I. H., BRINTON, L. A., LEVINE, P. H., CHEN, C. J. & YANG, C. S. (1999). Cigarette smoking, alcohol consumption and risk of nasopharyngeal carcinoma in Taiwan. *Cancer Cause Contr.* **10**, 201–7.

DIGGLE, P., KUNG-YEE LIANG, K. & ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.

FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.* **52**, 399–433.

GAIL, M. H., WIEAND, S. & PIANTADOSI, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regression and omitted covariates. *Biometrika* **71**, 431–44.

HILDESHEIM, A. & LEVINE, P. H. (1993). Etiology of nasopharyngeal carcinoma—a review. *Epidem. Rev.* **15**, 466–85.

HOUWING-DUISTERMAAT, J. J. & VAN HOUWELINGEN, H. C. (1998). Incorporation of family history in logistic regression models. *Statist. Med.* **17**, 2865–82.

NEUHAUS, J. M. (1993). Estimation efficiency and tests of covariate effects with clustered binary data. *Biometrics* **49**, 989–96.

NEUHAUS, J. M., KALBFLEISCH, J. D. & HAUCK, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int. Statist. Rev.* **59**, 25–35.

NEUHAUS, J. M., HAUCK, W. W. & KALBFLEISCH, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed effects logistic models. *Biometrika* **79**, 755–62.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

WITTE, J. S., GAUDERMAN, J. & THOMAS, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am. J. Epidem.* **149**, 693–705.

ZEGER, S. L., LIANG, K.-Y. & ALBERT, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–60.

[*Received May* 2000. *Revised January* 2001]